Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

CSE3506 Essentials of Data Analytics

Name : **Revanth Bhargava Boyidi**
Reg. No. : **19BCE1177**
Lab Exercise: **4-Time Series Forecasting**

Objective: To perform time series analysis on the "AirPassengers" dataset using R programming.

Methods:

i. Display entire dataset

ii. Check for unfilled data

iii. Display the statistical info of the dataset such as min, max, 1st quartile, 3rd quartile, mean and median.

iv. Plot 'data' (No. of Air passengers Vs Year)

v. Plot as timeseries 'data' (monthwise)

vi. Decompose the data as multiplicative and store as 'ddata'

vii. Plot 'ddata'

viii. Plot the following: trend, seasonal and random separately.

ix. Perform ADF test for stationarity

x. Plot ACF and PACF

xi. Model using ARIMA

xii. Conclusions

Store Airpassengers dataset (inbuilt dataset available in 'R') in a dataframe named "data", install packages such as 'forecast', 'tseries'.

```
setwd("D:\\SEM-VI\\EDA_CSE3506\\Lab\\Lab-4(11-02)_Time Series")
rm(list=ls())       #To clear the environment

data <- AirPassengers

#install.packages("r package", repos = "http://cran.us.r-project.org")
#install.packages('forecast')
```

```
#install.packages('tseries')
library('forecast')

## Warning: package 'forecast' was built under R version 4.1.2

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

library('tseries')

## Warning: package 'tseries' was built under R version 4.1.2
```

(i)Display entire dataset

```
data

##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 1949 112 118 132 129 121 135 148 148 136 119 104 118
## 1950 115 126 141 135 125 149 170 170 158 133 114 140
## 1951 145 150 178 163 172 178 199 199 184 162 146 166
## 1952 171 180 193 181 183 218 230 242 209 191 172 194
## 1953 196 196 236 235 229 243 264 272 237 211 180 201
## 1954 204 188 235 227 234 264 302 293 259 229 203 229
## 1955 242 233 267 269 270 315 364 347 312 274 237 278
## 1956 284 277 317 313 318 374 413 405 355 306 271 306
## 1957 315 301 356 348 355 422 465 467 404 347 305 336
## 1958 340 318 362 348 363 435 491 505 404 359 310 337
## 1959 360 342 406 396 420 472 548 559 463 407 362 405
## 1960 417 391 419 461 472 535 622 606 508 461 390 432
```

(ii)Check for unfilled data

```
length(data[is.na(data) == TRUE])

## [1] 0
```

Data doesn’t have unfilled data.

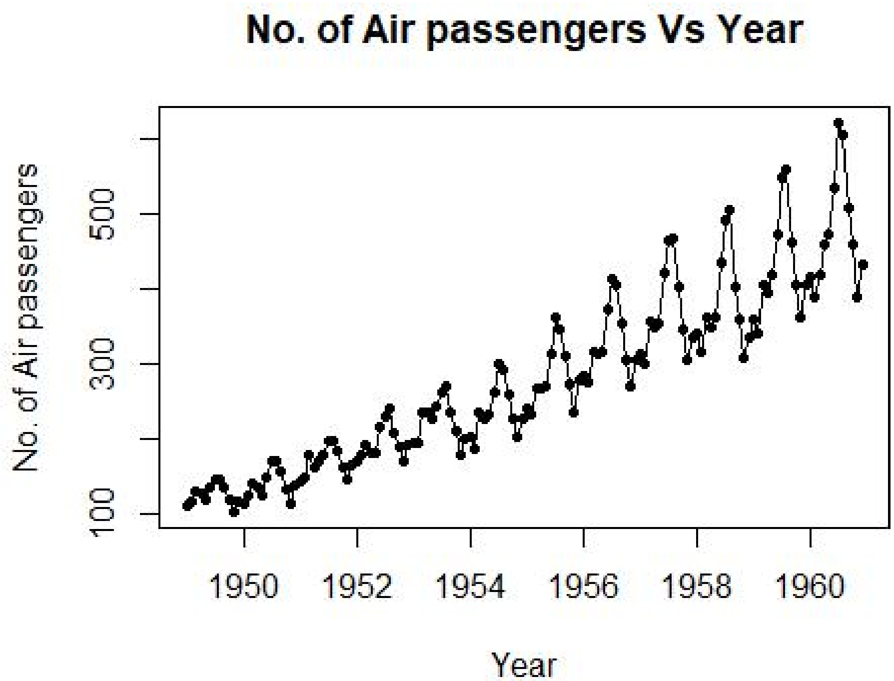
(iii)Display the statistical info of the dataset such as min, max, 1st quartile, 3rd quartile, mean and median.

```
summary(data)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   104.0   180.0   265.5   280.3   360.5   622.0
```

(iv)Plot ‘data’ (No. of Air passengers Vs Year)

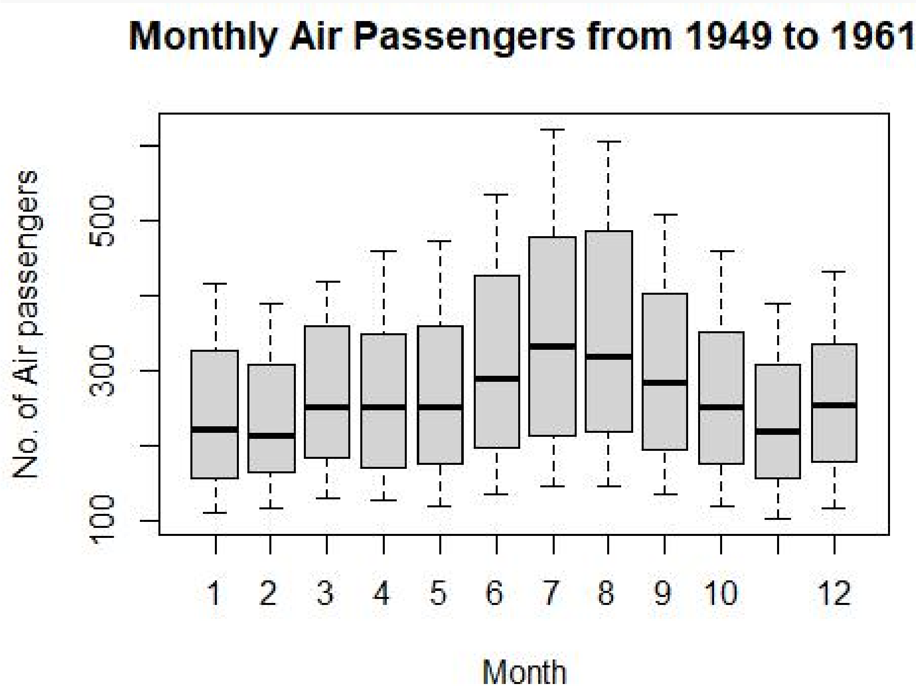
```
plot(data, xlab = 'Year', ylab = 'No. of Air passengers', main = 'No. o
f Air passengers Vs Year', type='o', pch=20)
```



It can be concluded from the graph that no. of passengers on an average were increasing every year.

(v)Plot as timeseries 'data' (monthwise)

```
boxplot(data~cycle(data), xlab = 'Month', ylab = 'No. of Air passengers
', main = 'Monthly Air Passengers from 1949 to 1961')
```



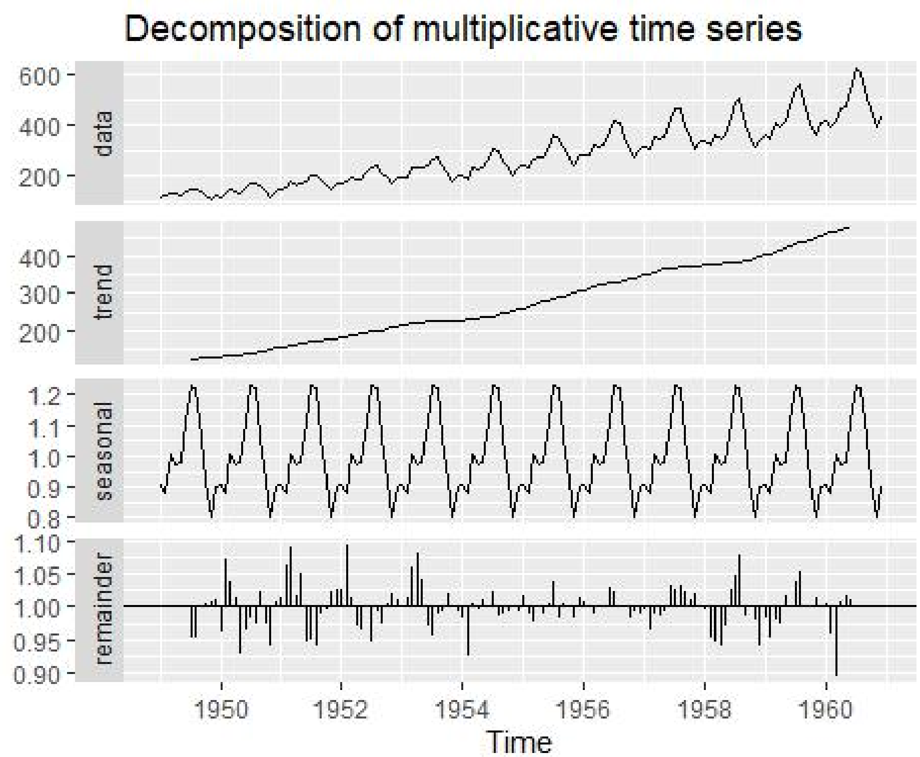
From months June - September, we can see a hike in number of passengers very year.

(vi)Decompose the data as multiplicative and store as 'ddata'

```
ddata <- decompose(data,"multiplicative")
```

(vii)Plot 'ddata'

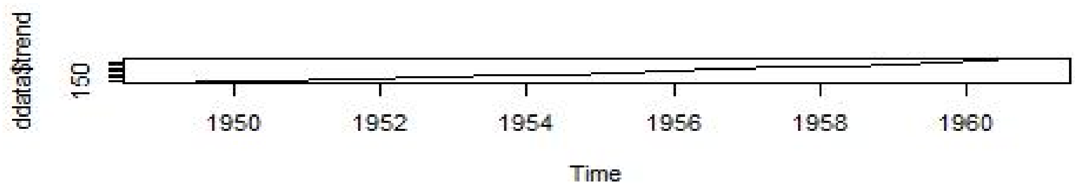
```
autoplot(ddata)
```

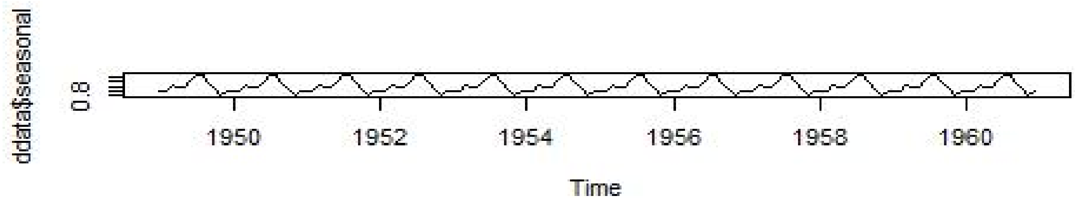


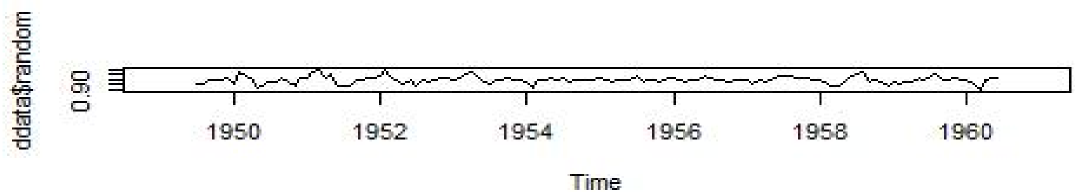
In these decomposed plots we can see the trend yearly and monthly. We can also observe the estimation of the random component depicted under the remainder.

(viii)Plot the following: trend, seasonal and random separately.

```
par(mfrow=c(3,1))
plot(ddata$trend)
plot(ddata$seasonal)
plot(ddata$random)
```







(ix)Perform ADF test for stationarity

```
adf.test(data, alternative="stationary", k=0)

## Warning in adf.test(data, alternative = "stationary", k = 0): p-valu
e smaller
## than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  data
## Dickey-Fuller = -4.6392, Lag order = 0, p-value = 0.01
## alternative hypothesis: stationary
```

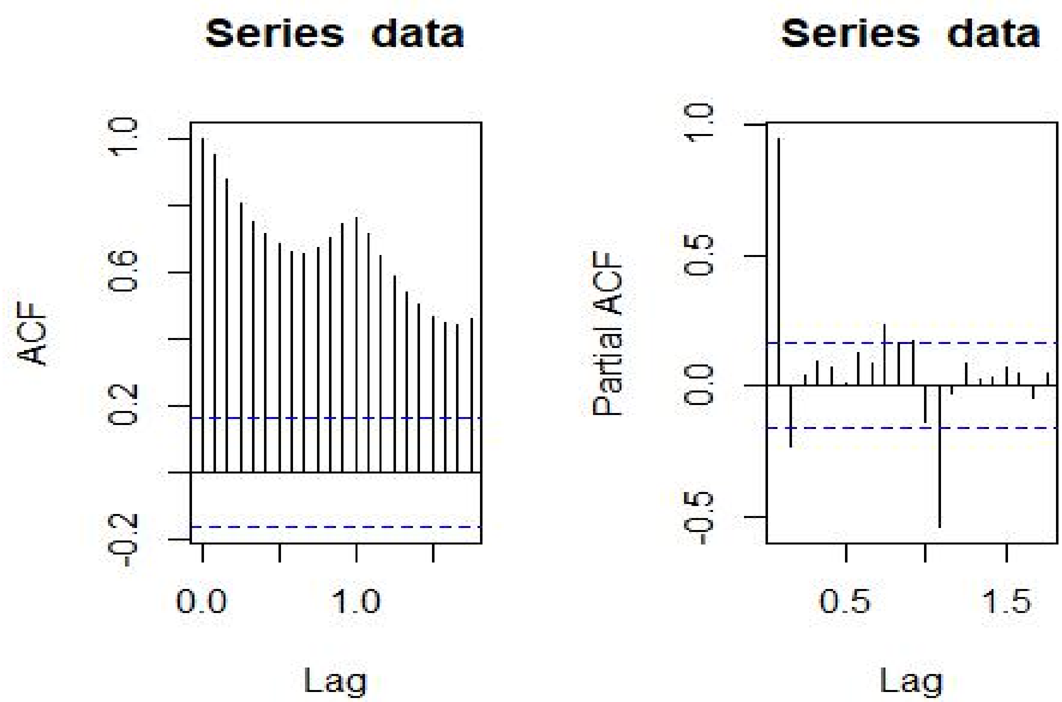
As per the test results above, the p-value is 0.01 which is <0.05 therefore we reject the null in favour of the alternative hypothesis that the time series is stationary.

(x)Plot ACF and PACF

ACF is an (complete) auto-correlation function which gives us values of auto-correlation of any series with its lagged values.

PACF is a partial auto-correlation function.

```
par(mfrow=c(1,2))
acf(data)
pacf(data)
```



(xi)Model using ARIMA

An auto-regressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. A statistical model is auto-regressive if it predicts future values based on past values.

```
auto.arima(data)

## Series: data
## ARIMA(2,1,1)(0,1,0)[12]
##
## Coefficients:
##          ar1     ar2      ma1
##       0.5960  0.2143  -0.9819
## s.e.  0.0888  0.0880   0.0292
##
## sigma^2 = 132.3:  log likelihood = -504.92
## AIC=1017.85   AICc=1018.17   BIC=1029.35
```

The ARIMA fitted model is:

$$\hat{Y} = 0.5960Y_{t-2} + 0.2143Y_{t-12} - 0.9819e_{t-1} + E$$

where E is some error.

Conclusions

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly.

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. ARIMA is an acronym that stands for Auto-Regressive Integrated Moving Average. It is a generalization of the simpler Auto-Regressive Moving Average and adds the notion of integration.

The ARIMA fitted model is: $\hat{Y} = 0.5960Y_{t-2} + 0.2143Y_{t-12} - 0.9819e_{t-1} + E$

From the above p-value, we can conclude that the residuals of our ARIMA prediction model is stationary.